WHAT IS CLAIMED IS:

1. In a switch having N_{in} input ports applied to K_{in} input shared blocks, a central switching fabric, and N_{out} output ports provided from K_{out} output shared blocks, a method for scheduling packets queued at the input shared blocks for application to the output ports, the method comprising steps of:

a) providing, for each of the input shared blocks, an indication of a number of links reserved by the input shared block to each of the output shared blocks;

b) providing each of the input shared blocks with a token, each token corresponding to an output shared block and including a value indicating a number of links available to the associated output shared block;

c) if it is determined that an input shared block needs links to an output shared block associated with a token held by the input shared block, then

i) reserving links, to the extent available as indicated by the token, to the output shared block

ii) updating the value indicating the number of links available to the associated output shared block, and

iii) updating the value of the indication of a number of links reserved by the input shared block to the associated output shared block.

2. The method of claim 1 further comprising a step of

d) passing the token to a next input shared block at the end of a reservation time slot.

3. The method of claim 2 wherein a cell slot includes at least one reservation slot and further comprising a step of:

d) delivering cells to the central switch fabric based on their currently reserved links at the end of each cell slot.

4. The method of claim 3 further comprising a step of:

e) destroying the tokens at the end of each cell slot; and

f) generating new tokens at each input shared block at the end of each cell slot.

5. In a switch having N_{in} input ports applied to K_{in} input shared blocks, a central switching fabric, and N_{out} output ports provided from K_{out} output shared blocks, a method for scheduling packets queued at the input shared blocks for application to the output ports, the method comprising steps of:

a) for each of the input shared blocks, providing a request token associated with one of the output shared blocks, each of the request tokens including an indication based on a number of requested links for the output shared block with which it is associated;

b) for each of the input shared blocks, providing a release token associated with one of the output shared blocks, each of the release tokens including an indication based on a number of released links for the output shared block with which it is associated;

c) for each of one or more reservation time slots within a cell time slot,

i) accepting, by an input shared block, a request token from another input shared block,

ii) determining whether a virtual output queue of the input shared block associated with the output shared block with which the request token is associated, is heavily occupied,

iii) if it is determined that the virtual output queue of the input shared block associated with the output shared block with which the request token is associated is heavily occupied, then

A) requesting at least one extra link to the output shared block associated with the accepted request token,

iv) determining whether the virtual output queue of the input shared block associated with the output shared block with which the request token is associated, is lightly occupied,

v) if it is determined that the virtual output queue of the input shared block associated with the output shared block with which the request token is associated is lightly occupied, then

A) releasing at least one link to the output shared block associated with the accepted request token if it is indicated that a number of requested links for the output shared block is greater than zero,

vi) if it is determined that the virtual output queue of the input shared block associated with the output shared block with which the request token is associated is not lightly occupied, then

A) releasing at least one link to the output shared block associated with the

accepted request token if the input shared block reserved more than a predetermined number of links and if it is indicated that a number of requested links for the output shared block is greater than zero,

vii) accepting, by the input shared block, a release token from another input shared block;

viii) determining whether or not the input shared block can take a link to the output shared block associated with the release token,

ix) if it is determined that the input shared block can take a link to the output shared block associated with the release token, then taking a link from the release token.

6. The method of claim 5 wherein the step of determining whether a virtual output queue of the input shared block associated with the output shared block with which the request token is associated, is heavily occupied, is based on a comparison with a threshold value.

7. The method of claim 5 wherein the step of requesting an extra link to the output shared block associated with the accepted request token, is effected by setting a request indicator corresponding to the input shared block and the output shared block, and incrementing the indication based on the number of links to the output shared block requested.

8. The method of claim 5 wherein the step of determining whether the virtual output queue of the input shared block associated with the output shared block with which the

request token is associated, is lightly occupied, is based on a comparison with a threshold value.

9. The method of claim 5 wherein the step of releasing a link to the output shared block associated with the accepted request token is effected by decreasing the indication based on the number of links to the output shared block released.

10. The method of claim 5 wherein the step of determining whether or not the input shared block can take a link to the output shared block associated with the release token, is based on a number of all reserved links by the input shared block and an indication of whether or not the input shared block had requested a link to the output shared block.

11. The method of claim 5 further comprising a step of:

d) delivering, by each of the input shared blocks, cells to the central switch fabric based on current indications of a number of link reservations to each of the output shared blocks, at the end of a cell time slot.

12. In a switch having N_{in} input ports applied to K_{in} input shared blocks, a central switching fabric, and N_{out} output ports provided from K_{out} output shared blocks, a method for scheduling packets queued at the input shared blocks for application to the output ports, the method comprising steps of:

a) for each of the input shared blocks, providing a request token associated with one of the output shared

blocks, each of the request tokens including an indication based on a number of requested links for the output shared block with which it is associated;

b) for each of the input shared blocks, providing a release token associated with one of the output shared blocks, each of the release tokens including an indication based on a number of released links for the output shared block with which it is associated;

c) for each of one or more reservation time slots within a cell time slot,

i) accepting, by an input shared block, a request token from an another input shared block,

ii) determining whether a virtual output queue of the input shared block associated with the output shared block with which the request token is associated, is heavily occupied,

iii) if it is determined that the virtual output queue of the input shared block associated with the output shared block with which the request token is associated is heavily occupied, then

A) requesting at least one extra link to the output shared block associated with the accepted request token,

iv) determining whether the virtual output queue of the input shared block associated with the output shared block with which the request token is associated, is lightly occupied,

v) if it is determined that the virtual output queue of the input shared block associated with the output shared block with which the request token is associated is lightly occupied, then

A) releasing at least one link to the output shared block associated with the accepted request token if it is indicated that a number of requested links for the output shared block is greater than zero,

vi) if it is determined that the virtual output queue of the input shared block associated with the output shared block with which the request token is associated is not lightly occupied, then

A) releasing at least one link to the output shared block associated with the accepted request token if the input shared block reserved more than a predetermined number of links,

vii) accepting, by the input shared block, a release token from another input shared block,

viii) determining whether to release at least one link to the output shared block with which the accepted release token is associated based on a queue occupancy, a number of links reserved, and a predetermined number of links

ix) if it has been determined to release a link to the output shared block with which the accepted release token is associated, releasing a link,

x) determining whether or not to take at least one released link to the output shared block with which the accepted release token is associated based on queue occupancy, a number of links reserved, and a number of links between the input shared block and the central switch fabric, and

xi) if it is determined to take at least one released link to the output shared block with which the accepted release token is associated, taking at least one link.

13. The method of claim 12 wherein the step of determining whether a virtual output queue of the input shared block associated with the output shared block with which the request token is associated, is heavily occupied, is based on a comparison with a threshold value.

14. The method of claim 12 wherein the step of requesting at least one extra link to the output shared block associated with the accepted request token, is effected by setting a request indicator corresponding to the input shared block and the output shared block, and incrementing the indication based on the number of links to the output shared block requested.

15. The method of claim 12 wherein the step of determining whether the virtual output queue of the input shared block associated with the output shared block with which the request token is associated, is lightly occupied, is based on a comparison with a threshold value.

16. The method of claim 12 wherein the step of releasing at least one link to the output shared block associated with the accepted request token is effected by decreasing the indication based on the number of links to the output shared block released.

17. The method of claim 12 wherein the step of determining whether or not the input shared block can take at least one link to the output shared block associated with the release token, is based on a queue occupancy of a virtual output queue, and a number of all reserved links by the input shared block and an indication of whether or not the input shared block had requested at least one link to the output shared block if it is indicated that a number of released links for the output shared block is greater than zero.

18. The method of claim 12 further comprising a step of:

d) delivering, by each of the input shared blocks, cells to the central switch fabric base based on current indications of a number of link reservations to each of the output shared blocks, at the end of a cell time slot.

19. A switch for switching packets arriving at a number of input ports to an appropriate one of a number of output ports, the switch comprising:

a) a central switching fabric;

b) output shared blocks, each coupled with at least one output port;

c) links between the central switch fabric and each of the output shared blocks

d) input shared blocks, each

i) coupled with at least one input port,

ii) having virtual output queues, each of the virtual output queues corresponding to one or more output ports,

iii) storing

A) an indication of whether at least one links to each of the output shared blocks has been requested,

B) an indication based on a number of links to each of the output shared blocks released, and

C) an indication based on a number of links reserved to each of the output shared blocks;

e) request tokens, each associated with a particular one of the output shared blocks and each indicating a number of requests for links to the associated one of the output shared blocks;

f) release tokens, each associated with a particular one of the output shared blocks and each indicating a number of released links to the associated one of the output shared blocks; and

g) links between the central switch fabric and each of the input shared blocks.

20. The switch of claim 19 wherein each of the input shared blocks holds at least one of the request tokens and at least one of the release tokens during a reservation time slot.